

Applied Regression Analysis

.....

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics, University of
Washington

Session 4

1

© 2002, 2003 Scott S. Emerson, M.D., Ph.D.

Correlation, Regression

.....

2

Pearson's Correlation Coefficient (r)...

- A measure of the tendency of the largest measurements for one variable to be associated with the largest measurements of the other variable
 - The sample correlation r estimates the population correlation ρ (rho)

3

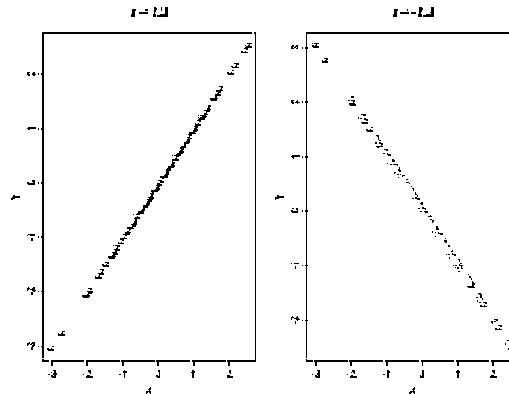
Pearson's Correlation Coefficient (r)...

- Range of r : $-1 \leq r \leq 1$
 - $r = 1$: perfect positive correlation
 - a graph of X vs Y will be a straight line with positive slope
 - $r = -1$: perfect negative correlation
 - a graph of X vs Y will be a straight line with negative slope
 - $r = 0$: no correlation

4

Pearson's Correlation Coefficient (r)...

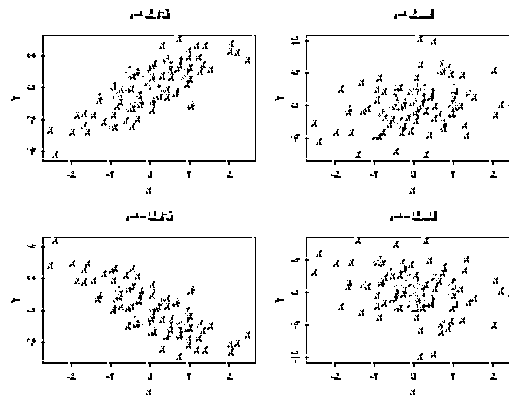
- Pearson's correlation coefficient with linear data



5

Pearson's Correlation Coefficient (r)...

- Pearson's correlation coefficient with variable data



6

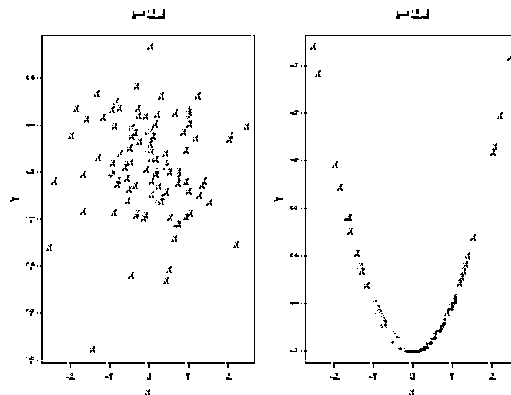
Pearson's Correlation Coefficient (r)...

- Correlation and Independence
 - Independent variables will have $\rho = 0$
 - (and r tending to be close to 0)
 - However, uncorrelated variables are not necessarily independent
 - Correlation is a measure of linear trend in the mean of one variable in groups defined by the other
 - It is possible that a nonlinear association exists between two variables, and that the first order trend is a zero slope

7

Pearson's Correlation Coefficient (r)...

- Pearson's correlation coefficient with nonlinear data



8

Pearson's Correlation: Stata Commands.....

-`"correlate varlist"`

- Correlation of all pairs of variables
- Missing data deleted on a casewise basis

-`"pwcorr varlist"`

- Correlation of all pairs of variables
- Missing data deleted on a pairwise basis

9

Example: Correlation in FEV Data.....

```
. corr subjid age fev height sex smoke
(obs=654)
```

	subjid	age	fev	height	sex	smoke
subjid	1.0000					
age	-0.0112	1.0000				
fev	-0.0147	0.7565	1.0000			
height	-0.0317	0.7919	0.8681	1.0000		
sex	0.0407	-0.0291	-0.2084	-0.1590	1.0000	
smoke	-0.0601	-0.4043	-0.2454	-0.2804	-0.0756	1.0000

- Some of these correlations don't make much sense

- subjid is a nominal variable
- sex, smoke are binary variables

10

Regression Setting

.....

11

Two Variable Setting

.....

- Many statistical problems can be regarded as considering the association between two variables
 - Response variable (outcome, dependent variable)
 - Grouping variable (predictor, independent variable)
 - The scientific question is addressed by comparing the distribution of the response variable across groups that are defined by the grouping variable
 - Within each group, the value of the grouping variable is constant

12

Correspondence to Number of Samples.....

- In introductory statistics courses, there is a tendency to characterize problems according to the number of samples and whether the samples are independent
 - The correspondence between that nomenclature and the two variable setting is based on the type of variable used as the grouping variable
 - Constant: One sample problem
 - Binary: Two sample problem
 - Categorical: k sample problem (e.g., ANOVA)
 - Continuous: Infinite sample problem
 - Regression

13

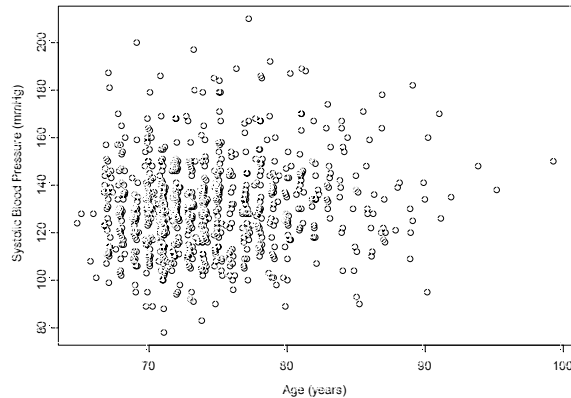
Infinite Sample Problem.....

- When the grouping variable is continuous, there are conceptually an infinite number of groups
 - E.g., when investigating the blood pressure across age groups
 - If measured with enough precision, no two people have exactly the same age
 - It is, of course, rare that we would have an infinite number of groups in our sample
 - (and possibly not even in our population)
 - It is common to have 1 (or fewer) subjects in a particular group in our sample

14

Example: SBP and Age

.....



15

Regression Methods

.....

- Regression can be thought of as extending one and two sample statistics (e.g., the t test) to the infinite sample problem
 - While we don't really ever have (or care) about an infinite number of samples, it is easiest to use models that would allow that in order to handle
 - Continuous predictors of interest
 - Adjustment for other variables

16

Regression vs Two Sample Methods.....

- A very convenient feature of the regression methods is that when used with a binary grouping variable they reduce to the corresponding two variable methods
 - Linear regression with a binary predictor
 - t test with equal variance
 - (approx t test with unequal variance when using “robust” standard errors)

17

Regression vs Two Sample Methods.....

- “Everything is regression.”
 - Scott Emerson

18

Linear Regression Setting: Example.....

- Association between blood pressure and age
 - Scientific question:
 - Does aging affect blood pressure?
 - Statistical question:
 - Does the distribution of blood pressure differ across age groups?
 - Acknowledges variability of response
 - Acknowledges uncertainty of cause and effect
 - » (Differences could be related to calendar time instead of age)

19

Linear Regression Setting Example.....

- Association between blood pressure and age (cont.)
 - Definition of variables
 - Response: Systolic blood pressure
 - continuous
 - Predictor of interest (grouping): Age
 - continuous
 - » an infinite number of ages are possible
 - » we probably will not sample every one of them

20

Linear Regression Setting: Example.....

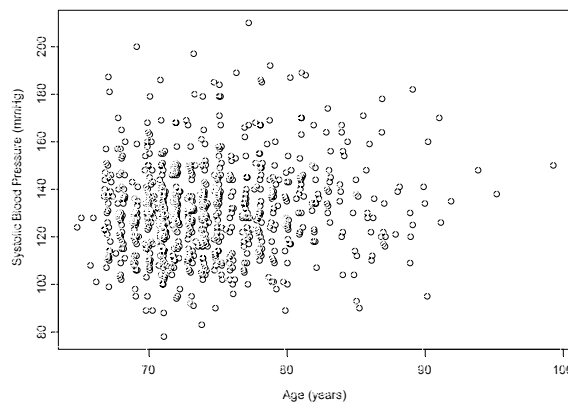
- Association between blood pressure and age (cont.)
 - Answering the question is possible if we try to assess linear trends in, say, average SBP by age
 - Estimate best fitting line to average SBP within age groups

$$E(SBP | Age) = \beta_0 + \beta_1 \times Age$$

- An association will exist if the slope (β_1) is nonzero
 - In that case, the average SBP will be different across different age groups

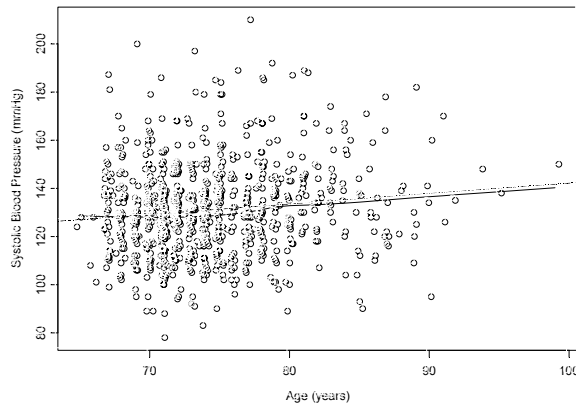
21

Linear Regression Setting: Example.....



22

Linear Regression Setting: Example.....



23

Linear Regression Setting Example.....

- Association between blood pressure and age (cont.)
 - The regression model thus produces something similar to “a rule of thumb”
$$E(SBP | Age) = 100 + 0.5 \times Age$$
 - E.g., “Normal SBP is 100 plus half your age”

24

Linear Regression Setting

Example.....

Actual estimates (and inference)

```
. regress sbp age
```

Source		SS	df	MS	Number of obs =	735
Model		4056	1	4056.4	F(1, 733) =	10.63
Residual		279740	733	381.6	Prob > F =	0.0012
Total		283796	734	386.6	R-squared =	0.0143
					Adj R-squared =	0.0129
					Root MSE =	19.536

sbp		Coef.	St.Err.	t	P> t	[95% Conf Int]
age		.431	.132	3.26	0.001	.172 .691
_cons		98.949	9.889	10.01	0.000	79.535 118.364

$$E(SBP | Age) = 98.9 + 0.43 \times Age$$

25

Linear Regression Setting

Example.....

- We can make inference about the regression estimates
 - The regression output provides
 - estimates
 - Intercept: estimated mean when age = 0
 - Slope: estimated difference in average SBP for two groups differing by one year in age
 - standard errors
 - confidence intervals
 - P values testing for
 - Intercept of zero (who cares?)
 - Slope of zero (test for linear trend in means)

26

Linear Regression Setting

Example.....

- In this example we are primarily interested in the slope
 - “From linear regression analysis, we estimate that for each year difference in age, the difference in mean SBP is 0.43 mmHg. A 95% CI suggests that this observation is not unusual if the true difference in mean SBP per year difference in age were between 0.17 and 0.69 mmHg. Because the P value is $P < .0005$, we reject the null hypothesis that there is no linear trend in the average SBP across age groups.”

27

Regression: Necessary

Ingredients.....

- Response variable
 - The distribution of this variable will be compared across the groups
 - Linear regression models the mean of this variable
 - Log transformation of the response corresponds to modeling the geometric mean
 - Notation:
 - It is extremely common (99 of 100 statisticians agree) to use Y to denote the response variable when discussing general methods

28

Regression: Necessary Ingredients.....

- Predictor (grouping) variables
 - Group membership is measured by a variable
 - Notation
 - When not using mnemonics, I will tend to use X to denote a predictor variable
 - (When we proceed to multiple regression, I will use subscripts to denote different predictors)

29

Regression: Necessary Ingredients.....

- Regression model
 - We typically consider a “linear predictor function” that is linear in the modeled predictors
 - Expected value (mean) of Y for a particular value of X

$$E(Y | X) = \beta_0 + \beta_1 \times X$$

- Interpretation of the “regression parameters”
 - Intercept β_0 : Mean Y for a group with $X=0$
 - Quite often not of scientific interest
 - » Often outside range of data, sometimes impossible
 - slope β_1 : Diff in mean Y for groups differing in X by 1 unit
 - Usually our measure of association between Y and X

30

Simple Linear Regression

- Simple linear regression of response Y on predictor X
 - Mean for an arbitrary group derived from model
 - Interpretation of parameters by considering special cases

Model	$E[Y_i X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x + 1$	$E[Y_i X_i = x + 1] = \beta_0 + \beta_1 \times x + \beta_1$

31

Simple Linear Regression

- Interpretation of the model
 - In simple linear regression, we assume that a graph of average response within a group (on Y axis) versus value of predictor within a group would be a straight line
 - Algebra: A line is of form $y = mx + b$
 - In the presence of variation of response within groups (i.e., in the real world), the line is describing the central tendency of the data in a scatterplot of the response versus the predictor

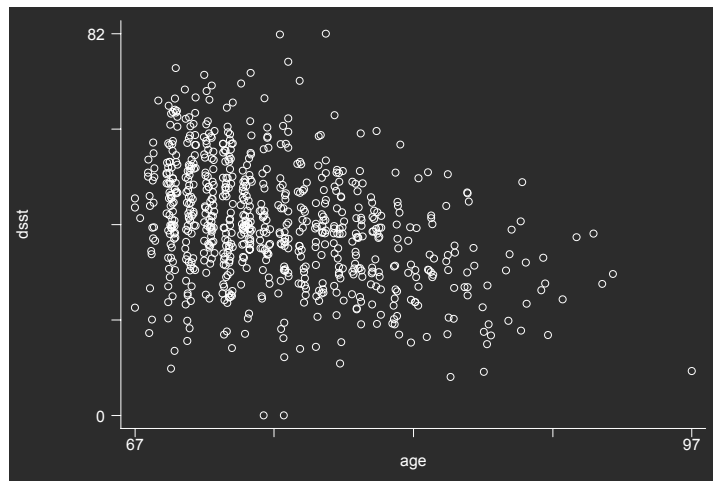
32

Simple Linear Regression: Example.....

- Trends in mental function with age
 - Cardiovascular Health Study
 - A cohort of ~5,000 elderly subjects in four communities followed with annual visits
 - Mental function measured at baseline by Digit Symbol Substitution Test (DSST)
 - Question: How does performance on DSST differ across age groups

33

Scatterplot of DSST versus AGE.....



34

Descriptives for DSST in Age Strata.....

Age	N	Nonmsgn	Mean	Std Dev
67	4	4	39.25	11.03
68	22	21	44.05	12.50
69	79	79	46.62	12.40
70	72	71	44.85	12.63
71	69	68	47.09	10.85
72	75	75	42.19	12.86
73	64	64	43.22	10.06
74	39	39	41.15	12.21
75	44	44	40.84	15.76
76	32	32	39.03	11.41
77	39	37	40.11	12.69

35

Descriptives for DSST in Age Strata.....

Age	N	Nonmsgn	Mean	Std Dev
78	36	36	38.56	11.11
79	33	33	36.61	9.78
80	28	28	36.21	8.90
81	19	19	32.95	11.84
82	15	14	30.93	8.94
83	12	12	35.08	9.06
84	14	12	29.92	12.18
85	9	9	35.56	9.37
86	7	7	18.43	5.71

36

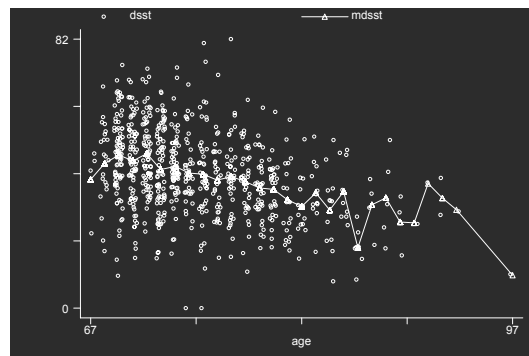
Descriptives for DSST in Age Strata

• Age	N	Nonmsgn	Mean	Std Dev
• 87	5	4	31.50	8.50
• 88	5	5	33.60	12.72
• 89	5	4	26.25	6.70
• 90	3	1	26.00	
• 91	1	1	38.00	
• 92	2	2	33.50	7.78
• 93	1	1	30.00	
• 97	1	1	10.00	

37

Plot of Mean DSST versus AGE

- sort age
- by age: egen mdsst = mean (dsst)
- graph dsst mdsst age, s(oT) c(.1) j(1)



38

Estimation of Least Squares Line

```
regress dsst age
(ANOVA table output omitted)
```

```
Number of obs =      723
F( 1, 721) = 116.81
Prob > F      = 0.0000
R-squared     = 0.1394
Adj R-squared = 0.1382
Root MSE     = 11.796
```

dsst	Coef.	StErr	t	P> t	[95% CI]	
age	-.938	.087	-10.81	0.000	-1.11	-.768
_cons	111.	6.48	17.11	0.000	98.25	124.

39

Interpretation of Stata Output

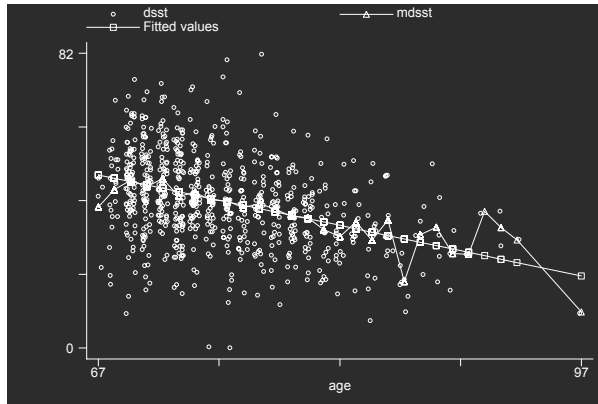
- Estimates of regression parameters
 - Intercept is labeled “_cons”
 - Estimated intercept: 111.
 - Slope is labeled by variable name: “age”
 - Estimated slope: -.938
 - Estimated linear relationship:
 - Average DSST by age given by

$$E[DSST_i | Age_i] = 111 - 0.938 \times Age_i$$

40

Superimposed Plot of Least Squares Line.....

- `predict fdsst`
- `graph dsst mdsst fdsst age, s(ots) c(.11) j(1)`



41

Interpretation of Stata Output.....

- Scientific interpretation of the intercept

$$E[DSST_i | Age_i] = 111 - 0.938 \times Age_i$$

- Estimated mean DSST for newborns is 111
 - Pretty ridiculous estimate
 - We never sampled anyone less than 67
 - Maximum value for DSST is 100
 - Newborns would in fact (rather deterministically) score 0
- In this problem, the intercept is just a mathematical construct to fit a line over the range of our data

42

Interpretation of Stata Output

.....

- Scientific interpretation of the slope

$$E[DSST_i | Age_i] = 111 - 0.938 \times Age_i$$

- Estimated difference in mean DSST for two groups differing by one year in age is -0.938, with older group averaging a lower score
 - For 5 year age difference: $5 \times -0.938 = -4.69$
 - For 10 year age difference: -9.38
- (If a straight line relationship is not true, we can still interpret the slope as an average difference in mean DSST per one year difference in age)

43

Interpretation of Stata Output

.....

- Comments on scientific interpretation of the slope
 - Note that I express this as a difference between group means rather than a change with aging
 - We did not do a longitudinal study
 - To the extent that the true group means have a linear relationship, this interpretation applies exactly
 - If the true relationship is nonlinear
 - The slope estimates the “first order trend” for the sampled age distribution
 - We should not regard the estimates of individual group means as accurate

44

Alternative Representation of Model.....

- Sometimes linear regression models are expressed in terms of the response instead of the mean response

$$\text{Model} \quad Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i$$

- The response is divided into two parts
 - The mean (systematic part or “signal”)
 - The “error” (random part or “noise”)
 - difference between the observed value and the corresponding group mean
 - ε_i is called the error
- The error distribution describes the within-group distribution of response

45

Interpretation of Stata Output.....

- Estimates for error distribution
 - The error distribution is estimated from the residuals

$$\text{Residual} \quad \hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \times X_i)$$

- The mean of the errors is assumed to be 0
- The sample standard deviation of the residuals is reported as the “Root Mean Squared Error”
- Thus we estimate within group SD of 11.796 in the DSST vs age example
 - Classical linear regression: SD for each age group
 - Robust standard error estimates: An average across groups

46

Inferential Uses of Regression Models

- Regression models can be used to answer the most commonly encountered statistical questions
 - Prediction
 - Estimating a future observation of response Y
 - Often we use the mean or geometric mean
 - Quantifying distributions
 - Describing the distribution of response Y within groups by estimating the mean $E(Y | X)$
 - Comparing distributions across groups
 - Distributions differ across groups if the regression slope parameter β_1 is nonzero

$$E(Y | X) = \beta_0 + \beta_1 X$$

47

Statistical Validity of Inference

- Inference (CI, P values) about associations is based on two general types of assumptions
 - Assumptions about independence of observations
 - Classically: All observations are independent
 - Robust standard error estimates: Allow correlated observations within identified clusters
 - Assumptions about variance of observations within groups
 - Classically: Equal variances across groups
 - Robust standard error estimates: Allow unequal variances across groups

48

Statistical Validity of Inference

.....

- Inference (CI, P values) about mean response in specific groups has a further assumption
 - Assumption about adequacy of linear model
 - Classically OR robust standard error estimates:
The mean response in groups is linear in the modeled predictor
 - (We can model transformations of the measured predictor)

49

Statistical Validity of Inference

.....

- Inference (prediction intervals, P values) about individual observations in specific groups has still another assumption
 - Assumption about distribution of errors within each group
 - Classically: The distribution of errors follows the same normal distribution within each group
 - Possible extension: The distribution of errors follows the same distribution within each group, though it need not be normal
 - This extension is not implemented in any software that I know of
 - (Inappropriate inference if robust standard error estimates are necessary for unequal variances)

50

Inference About Associations

- Inference about associations is far more robust than estimation of group means or individual predictions
 - If the response and predictor of interest were totally independent, the mean response in each group would be the same
 - A flat line would describe the mean response across groups (and a linear model is correct)
 - A nonzero slope suggests the presence of an association between mean response and predictor
 - The assumption of straight line relationships in the modeled (transformed) parameter need not hold exactly for examining such associations
 - (I am not modeling the data; instead I am testing for trends in the parameter – looking at “contrasts”)

51

Interpreting Inference for Association.....

- Robust interpretation of “positive” studies (statistically significant nonzero slopes)
 - “Statistically significant slope”
 - The observed data is atypical of a setting in which the mean response is the same across all groups
 - Data suggests evidence of a trend toward larger (smaller) means in groups having larger values of the predictor
 - The slope estimate (and CI) describe some sort of an average trend over the distribution of predictors in the sample
 - (Only if a straight line is a good description of the trend in the parameters can we also use the model to predict the mean or individual observations for each group)

52

Interpreting Inference for Association.....

- When using regression to detect associations, the interpretation of “lack of statistical significance” must take into account all possibilities
 - There may be no association
 - There may be an association but not in the parameter considered (i.e, the mean response)
 - There may be an association in the parameter considered, but the best fitting line has a zero slope (a curvilinear association in the parameter)
 - There may be a first order trend in the parameter, but we lacked statistical precision to be confident that it truly exists (type II error)

53

Regression Inference in Stata.....

- Stata allows inference based on either classical linear regression or robust standard error estimates
 - Classical linear regression
 - `regress respvar predictor`
 - E.g., `regress dsst age`
 - Robust standard error estimates
 - `regress respvar predictor, robust`
 - E.g., `regress dsst age, robust`
 - The two approaches differ in CI and P values, not estimates

54

Interpretation of Stata Output

- Inference with regression models
 - Linear regression intercept and slope parameters are asymptotically normally distributed, thus all we need to know in addition to the estimate (and interpretation) is the standard error
 - Stata automatically provides
 - standard error estimates
 - two-sided P values of a test that the regression parameters are 0
 - 95% confidence intervals
 - and a lot of other statistics, most of which (to my mind) are unnecessary to see

55

Classical Linear Regression

```
regress dsst age
Number of obs =      723
F( 1, 721) = 116.81
Prob > F      = 0.0000
R-squared     = 0.1394
Adj R-squared = 0.1382
Root MSE     = 11.796
```

dsst	Coef.	StErr	t	P> t	[95% CI]	
age	-.938	.087	-10.81	0.000	-1.11	-.768
_cons	111.	6.48	17.11	0.000	98.25	124.

56

Classical Linear Regression

- Inference about an association based on slope
 - Estimated trend in mean DSST by age is an average difference of -.938 for one year differences in age
 - T statistic: -10.81 (Who cares?)
 - P value: < .0001
 - CI for trend: -1.11, -.768
 - Conclusion: This is not what we would expect to see when no association exists in mean DSST by age

57

Robust Standard Error Estimates

```
regress dsst age, robust
Number of obs =      723
F( 1, 721) = 134.35
Prob > F      = 0.0000
R-squared     = 0.1394
Root MSE     = 11.796
```

dsst	Robust		t	P> t	[95% CI]	
	Coef.	StErr				
age	-.938	.081	-11.59	0.000	-1.10	-.779
_cons	111.	6.10	18.19	0.000	99.00	123.

58

Robust Standard Error Estimates.....

- Inference about an association based on slope
 - Estimated trend in mean DSST by age is an average difference of -.938 for one year differences in age
 - T statistic: -11.59 (Who cares?)
 - P value: < .0001
 - CI for trend: -1.10, -0.779
 - Conclusion: This is not what we would expect to see when no association exists in mean DSST by age

59

Choice of Inference Using Regression.....

- Which inference is correct?
 - Classical linear regression and robust standard error estimates differ in the strength of necessary assumptions
 - As a rule, if all the assumptions of classical linear regression hold, it will be more precise
 - (Hence, we will have greatest precision to detect associations if the linear model is correct)
 - The robust standard error estimates are, however, valid for detection of associations even in those instances

60

Choosing the Correct Model

.....

- “All models are false, some models are useful.”
 - George Box

61

Choosing the Correct Model

.....

- “In statistics, as in art, never fall in love with your model.”
 - Unknown

62

Model Checking

.....

- Much statistical literature has been devoted to means of checking the assumptions for regression models
 - I believe model checking is generally fraught with peril, as it necessarily involves multiple comparisons

63

Model Checking

.....

- “Blood suckers hide ‘neath my bed”
 - “Eyepennies”, Mark Linkous (Sparklehorse)

64

Model Checking

.....

- We cannot reliably use the sampled data to assess whether it accurately portrays the population
 - We are worried about what data we might not have seen
 - It is not so much the monsters that we see that scare us, but the goblins in the closet
 - (But we do worry more when we see a tendency to outliers in the sample or clear departures from the model)

65

Choice of Inference Using Regression.....

- My general recommendation: There is relatively little to be lost and much accuracy to be gained in using the robust standard error estimates
 - Avoids the need for “model checking”
 - (And “model checking” has too large an element of data driven analysis for my taste)
 - More logical scientific approach
 - Minimizes the need for assumptions that presume more detailed knowledge than the question we are trying to answer
 - E.g., if we don’t know how means might differ, why presume that we know how variances and higher moments behave?

66

Choice of Inference Using Regression.....

- Inference about estimation of group means or individual predictions should be interpreted extremely cautiously
 - The dependence on knowing the correct model and distribution means that we cannot be as confident in the estimates and inference
 - Nevertheless, such estimates are often the best approximations
 - Interpolation to unobserved groups is less risky than extrapolation outside the range of predictors